INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

CENTRO DE INVESTIGACIÓN ECONÓMICA

Discussion Paper Series

The Catalan Premium: Language and Employment in Catalonia

Silvio Rendón

Instituto Tecnológico Autónomo de México

May 2006

Discussion Paper 06-04

Av. Camino a Santa Teresa # 930
Col. Héroes de Padierna
México, D.F. 10700
M E X I C O

The Catalan Premium:

Language and Employment in Catalonia

Sílvio Rendon*

Centro de Investigación Económica - Instituto Tecnológico Autónomo de México (ITAM)
Av. Camino Santa Teresa 930, 10700, Mexico D.F., Mexico
Fax: +52 55 5628 4058; Email: srendon@itam.mx

Abstract.- In this paper I measure the contribution of knowing Catalan to finding a job in Catalonia. In the early eighties a drastic language policy change (*normalització*) promoted the learning and use of Catalan in Catalonia and managed to reverse the falling trend of its relative use versus Castilian (Spanish). Using census data for 1991 and 1996, I estimate a significant positive Catalan premium: the probability of being employed increases between 3 and 5 percentage points if individuals know how to read and speak Catalan; it increases between 2 and 6 percentage points for writing Catalan.

JEL Classification: J61, J70, J31, I20.
Keywords: Language, Immigration, Skill Premium.

*I thank Manuel Arellano, Jordi Galí, Albert Marcet, Núria Quella, Ernesto Villanueva, and participants at the 2004 ESPE Meetings in Bergen and of seminars at Universitat Pompeu Fabra, Universidad Carlos III, CentrA, CEMFI, and FEDEA for helpful discussions. I also thank the Institut d'Estadística de Catalunya (IDESCAT) for providing the data and Christian Dustmann and two anonymous referees for useful comments and suggestions. Financial support of the Spanish Ministry of Science and Technology (Grant SEC 2001-0674) and the Mexican Association of Culture is gratefully acknowledged. Marta González provided excellent research assistance in an early stage of this project. The usual disclaimer applies.

1 Introduction

What is the economic value of knowing Catalan in Catalonia for an individual? During most of the twentieth century Catalan, although the official language for most of Catalonia's history, was given an informal status, thus losing part of its economic value. Public policies imposed the learning and use of Castilian (Spanish[1]) over Catalan until the mid-seventies; however, they were reversed at the beginning of the eighties, and the learning and use of Catalan have since increased. This switch in language policy (known as *normalització*) implied not only that *immigrants* had to learn Catalan, the local language, to assimilate to a new economy, but also that natives had to become more proficient in the use of their own language. It is clear that in current times Catalan has increased its economic value; however, it is not clear how much this value represents.

In this article, I measure the contribution of knowing Catalan to the probability of being employed in Catalonia. This research adds to the rapidly growing literature on the returns to language proficiency in multilingual labor markets, such as Switzerland (Cattaneo and Winkelmann 2005), Wales (Drinkwater and O'Leary 1997, Henley and Jones 2003, Clark and Drinkwater 2002), Canada (Christofides and Swidinsky 2002), or Finland (Saarela and Finnäs 2003) among others. I use a sample of census data for the last two available census years, 1991 and 1996, on socioeconomic variables, employment status, and knowledge of the Catalan language. Accounting for the determinants of selection into knowing Catalan and verifying the sensitivity of results to the specification used, I estimate the probability of being employed conditional on the level of Catalan knowledge by gender and census year. Speaking *and* reading Catalan increases the probability of being employed by between 3 and 5 percentage points, whereas writing Catalan increases the probability of being employed by between 2 and 6 percentage points. This premium is higher for women than for men and positive both for individuals who know Catalan and for those who do not know it. Beyond this economic reward, I find that non-economic considerations are also

very important in explaining Catalan proficiency. In particular, externalities, such as living in a community where people know Catalan and are Catalan born, increase the probability that an individual knows Catalan.

As Chiswick and Miller (1995) explain, language has all the characteristics of human capital: it is embodied in the person, it is productive in the labor market and/or in consumption, and it is generated at a sacrifice of time and out-of-pocket resources. Most of the literature on language has focused on measuring the effect of its acquisition on the assimilation of immigrants in several recipient economies: USA, UK, Canada, Australia, Israel, and Germany (McManus 1985, Borjas 1987, Kossoudji 1988, Dustmann 1994, Rivera-Batiz 1990, Chiswick and Miller 2002, Dustmann and Fabbri 2003). Language, however, has an economic value also for the native population. Most economies are composed of different speech communities, originated by immigration or not, which communicate using related but different verbal and nonverbal languages (Lang 1986), and that may converge toward one only language (Lazear 1999). In this line, the official language may change. For example, Angrist and Lavy (1997) study the effect of arabization or the switch from French to Arabic instruction, in the local population of Morocco, and find that there was a decrease in the returns to schooling. Other cases in Europe are: from German to Hungarian in Hungary, or from Russian to Estonian in Estonia and to their national languages in other Baltic countries. As with these language changes, the change experienced in Catalonia in the eighties involved not only the 'immigrants' but also the 'natives.'

[Figure 1 here]

The evolution of the percentage of the population that speaks Catalan in Catalonia, Valencia, and the Balearics[2] is portrayed in Figure 1. This percentage has been steadily decreasing in Valencia and the Balearics since 1930, but in Catalonia it started increasing after 1975, the years of the transition to democracy in Spain, and

accelerated in the mid eighties reaching in 1996 the maximum level it used to have in the thirties. This acceleration appears to be the result of the Catalan government's public policies oriented to promote the learning and the use of Catalan. These policies were either absent or weaker in Valencia and in the Balearics.

Because of the coexistence of Castilian and Catalan, Catalonia is commonly perceived to be a bilingual society. However, several sociolinguistic studies show that Catalonia is characterized by diglossia rather than bilingualism (Vallverdú 1970, Strubell i Trueta 1981, Tudela 1986, Woolard 1989). Bilingualism means that two languages coexist and are used regularly and indistinctly for similar functions and with equal level of proficiency. On the contrary, diglossia happens when one language is reserved for education, formal communication and literature and the second or diglossic language is used for family and ordinary life, popular culture and folklore.

The economic value of a diglossic language, expressed in the wage or employment probability premium, is not expected to be very large. Given that the consensual objective in Catalonia has been to eliminate diglossia and move on to bilingualism, it is not surprising to see that active language policies have contributed to increase the economic value of knowing and using Catalan in Catalonia. This paper is the first attempt to measure the economic value of knowing Catalan for an individual. Currently, limitations on the availability of data only allow computation of this economic value in terms of employment rates and not in terms of wages, as it is usually done. Only when and if future employment and wage surveys in Catalonia include information on language variables will this become possible.

This paper is organized as follows. Section 2, reviews language policy in Catalonia. Section 3 discusses the data used in the estimation. Section 4 presents the results: the estimated coefficients, a discussion on identification of the premium, the predicted and counterfactual Catalan knowledge and employment probabilities, the premium, and a comparison of these results under different specifications. Section 5 summarizes the main conclusions of this paper.

2 Language Policy in Catalonia

The main official languages in Spain are Castilian, Catalan, Basque and Galician. Castilian, Catalan, and Galician, unlike Basque, are members of the Romance language family.[3] Presently, Castilian is official in the totality of the territory, and each of the other three languages shares co-officiality with Castilian in its own territory, altogether holding forty percent of the population of Spain. After the Spanish Civil War (1936-39) and up to the death of Franco in 1975, all public use of any language other than Castilian was abolished; and, until the Spanish Constitution was passed in 1978, the only official language anywhere in Spain was Castilian.[4]

In 1980 the Generalitat, the autonomous government of Catalonia, was re-established and gradually recovered its pre-war competencies. At that time, the massive immigration flows from the rest of Spain received in the sixties, together with the decades-long absence of all public and formal use of Catalan, had allowed diglossia to institutionalize. Moreover, high proportions of the generations schooled in the period from 1939 to 1975 were not able to read or write in Catalan, even if it was their native language. Thus, at the beginning of the eighties the Generalitat actively promoted Catalan becoming the normal language of public and private use in Catalonia. This is known as the "Normalization policy" (*normalització*), a law passed unanimously in the Catalan Parliament that extended the use of Catalan in the fields of education, public administration and public media. During this period Castilian was progressively replaced by Catalan as the main language of instruction in primary, middle and secondary schools.[5] In 1998, the Catalan Parliament approved a new law, which extends the presence of Catalan to previously unregulated domains, such as privately owned media, cultural industries and the socioeconomic sphere. These public policies, absent or weaker in other Catalan-speaking areas, seem to have significantly contributed to the recovery of the Catalan language in Catalonia depicted in Figure 1.

3 Data

Data come from the Catalan and Spanish National Statistical Institutes (IDESCAT-INE). I use two samples of 250,000 randomly selected individuals each from census data for years 1991 and 1996. Available variables are personal attributes such as gender, age, marital status, schooling, place of residence, place of birth, number of years in Catalonia, occupational status and knowledge of Catalan.[6] I also use data at the district area level called *municipi*, which capture the externality effects of the community on individuals' employment status and Catalan knowledge.

I select the main household members, that is, only parents and children aged between 16 and 60, born in Spain[7] and who are in the labor force (a condition which is particularly restrictive for women, as their participation rate is lower than men's). The selected sample consists of 96,863 individuals for year 1991, and of 96,985 for 1996. Appendix A.1 gives further details on the sample selection.

Respondents declare whether they understand, read, speak and write Catalan. Because respondents, guided by the proximity between Catalan and Castilian, may overreport their knowledge of Catalan,[8] I use a stringent definition of language knowledge. I class individuals who claim to understand, plus either speak or only read Catalan as having a basic level. Reading *and* speaking Catalan is the intermediate level, while writing is the superior level. Appendix A.2. explains the definition of the remaining variables in more detail.

[Table 1 here]

Table 1 presents descriptive statistics by subsample and by variable and reveals that Catalan knowledge for all groups increases from 1991 to 1996; in both years women are more proficient in Catalan than men. By contrast, employment rates decrease for all groups from 1991 to 1996. Individuals not born in Catalonia exhibit substantially lower levels of Catalan knowledge but their employment rates are not

substantially different from the average.[9] Additionally, employment rates are increasing in Catalan knowledge, except for men who write.

Women are younger and more proficient in Catalan, have more schooling and lower marriage rates than men. For both genders marriage rates decrease over time, whereas years of schooling increase over time.[10] Most of the population of Catalonia resides in the province of Barcelona. Approximately one third of the sample consists of people born outside Catalonia; around one third of them arrived when they were no older than 10, more than two thirds come from Andalusia, and very few from Valencia, the Balearics or La Franja (a thin stretch in neighboring Aragon), areas where Catalan also shares co-officiality with Castilian. The average number of years since migration is between 23 and 25 in 1991, that is, on average immigrants arrived to Catalonia in the second half of the sixties.

[Table 2 here]

As a starting point, it is instructive to estimate a simple probit model with a dummy for Catalan knowledge. Table 2 shows the coefficients for reading and speaking, and writing Catalan: they are positive and significant, and higher for women than for men. However, these results may just reflect that more employable individuals happen to be more proficient in Catalan. If so, we still do not know whether Catalan knowledge increases the probability of being employed. In the next section, I compute the premium correcting for possible selection effects.

4 Estimation

As Willis and Rosen (1979), I proceed in two stages. In the first stage selection into Catalan knowledge depends on economic and non-economic factors (knowing a language implies an economic reward and a satisfaction, conditional on contextual

factors like family background and language most used in the community). In the second stage, only economic rewards account for employment outcomes.[11]

4.1 Catalan Knowledge

Table 3 reports the coefficients of the selection equation into reading and speaking Catalan on the left side and into writing Catalan on the right side. The covariates used in the estimations are the variables discussed in the previous section, including square and interaction terms.[12] There are some clear common features in the estimated coefficients for the four subsamples. Individuals with more years of schooling are more likely to know Catalan. Older individuals are more likely to read and speak Catalan, but less likely to write it, possibly a result of being schooled during the years in which Catalan was diglossic. They will actively use the language in informal communication but not officially or in writing. The probability of knowing Catalan is thus increasing both in schooling and age, but at a decreasing rate. Older individuals exhibit lower returns to schooling, especially in the probability of writing Catalan.

[Table 3 here]

Individuals affected by Normalization are more likely to read and speak Catalan. This effect is weaker in 1996 than in 1991, greater for men than for women, and greater for reading and speaking than for writing skills.

Individuals who are not married or live outside Barcelona are more likely to know Catalan, especially to read and speak it. In areas with a higher percentage of employment in services, arguably the more urban areas, it is less likely that an individual knows Catalan. On the contrary, in areas with a higher percentage of individuals who know Catalan, it is more likely that an individuals knows Catalan, which is indicative of the importance of social interactions at the local level. In areas where the percentages of people being employed or being born in Catalonia are higher an individual is

more likely to read and speak Catalan, but less likely to write it. As with age, this difference across Catalan skills may stem from growing up in the times when Catalan was a diglossic language.

Individuals who were not born in Catalonia are less likely to know Catalan; however, individuals who arrived to Catalonia when they were younger than 10 are more likely to know Catalan. More exposure to the local culture, captured by the variable *years since migration*, makes language assimilation more likely. However, the effect of local exposure on knowing Catalan is stronger for individuals who arrived at a mature age.[13] Individuals who were born in Andalusia are less likely to know Catalan, especially women. On the contrary, individuals born outside Catalonia in areas were Catalan is used, such as Valencia, Balearics and La Franja, are more likely to know Catalan.

I also report indicators of overall fit, pseudo R^2 statistics. They are around 0.30, a relatively high value for a discrete variable, which reveals the fairly good explanatory power of the covariates in this estimation.

With the estimated coefficients I compute the predicted probabilities of knowing Catalan by skill level and perform two counterfactual experiments: for each sample I reestimate the language probability switching parameters by gender and census year. These counterfactuals are equivalent to decomposing the differences across genders and over time into differences caused by different individual attributes and different estimated coefficients. The results, reported in Table 4, show that women are more likely to read and speak and, especially, to write Catalan than men, and, unlike men, they experiment a significant increase in the probability of knowing Catalan from 1991 to 1996. These differences in gender and over time are substantially smaller than those suggested by a first inspection of the descriptive statistics in Table 1.

[Table 4 here]

4.2 Identification

Identification of a recursive bivariate probit model requires exogenous sources of variation (Maddala 1983). Otherwise, unmeasured individual characteristics that are correlated with employment may produce biased results. Even in this non-linear limited-dependent variable estimation, the effect of the unaccounted language variable on employment may be, in fact, containing the effect of other variables that determine Catalan knowledge. To reinforce identification of this effect in the employment equation one has to exclude variables used in the selection equation. The instruments I use are municipal variables capturing the externality effect of the community of residence in Catalan knowledge, origin variables, variables relevant for individuals born outside Catalonia (years since migration and age of arrival), and a variable indicating whether the individual was affected by Normalization.

Including origin variables as instruments may at first seem problematic. Generally speaking, individuals *born abroad*, immigrants, differ from indigenous people in more than their knowledge of the local language. However, Catalonia is a very particular scenario, and more than one of the generally accepted assumptions do not hold:

1. there is no physical (racial), religious, or legal difference between Spanish citizens born in Catalonia and those born in the rest of Spain; individuals not born in Catalonia are not unassimilated newcomers: as we have seen in Section 3, they have spent, on average, between 23 and 25 years in Catalonia, and one third of them arrived before they were 10 years of age. Their children were born in Catalonia and, especially if affected by Normalization, are fluent in Catalan. Moreover, and regardless of their level of proficiency in the Catalan language, these individuals identify themselves as Catalans and are considered Catalans by those born in Catalonia;

2. origin does not seem to play a role in accounting for employment status. Even at the descriptive level the employment rates of individuals not born in Catalonia are not substantially different from the average (see Section 3). It was precisely the availability of jobs that brought them to Catalonia in the first place;[14]

3. origin does account to *some* extent for the knowledge of Catalan. All individuals in Catalonia know Castilian, but not all of them, born or not in Catalonia, know Catalan.[15] As shown in Sub-section 4.1, individuals not born in Catalonia tend to be less proficient in the Catalan language than individuals born in Catalonia. However, older individuals, born or not in Catalonia, tend to be fluent in reading

and speaking Catalan, but not in writing it. This may be the result of the lack of formal education in Catalan when it was not official: knowing a diglossic language such as Catalan was not an asset.

4. finally, as explained in Section 3, to minimize the measurement error in Catalan knowledge caused by self-report, I used a stringent definition of language proficiency.

These arguments are illustrative of the pertinence of including regional origin variables as instruments in estimating the employment equation for this scenario: if the reasons explained above hold, these instruments are not correlated with employment status, but they are with the knowledge of Catalan. However, a few words of caution about results based on this identification strategy are in order. As shown in the next sub-section, it is straightforward to show that regional origin variables are powerful in explaining Catalan knowledge. It is, however, less obvious to determine formally the independence between regional origin variables and employment status: there are no established tests of instruments for limited dependent variable models. For example, although people born in Catalonia exhibit similar and slightly lower employment rates than those born outside Catalonia, a concern could be that the former may have better networks than the latter and end up with better jobs. This consideration would weaken the independence assumption and thereby the identification of the language premium. Thus, the reader should take results cautiously and bear in mind that these results are subject to improvement when richer datasets and wider sets of instruments become available in future research. I discuss this issue further in the next sub-sections.

4.3 Employment by Catalan Knowledge

The second stage consists of estimating the employment probability conditional on a given Catalan proficiency level, using the estimated parameters of the selection equation. These results are reported in Table 5, which reveals that individuals with more schooling or older are more likely to be employed; at the same time, there are decreasing returns for schooling and age, but no significant interaction between these

two variables. On the other hand, married men have significantly higher employment probabilities than single ones; by contrast married women have lower employment probabilities than single ones. Individuals who reside outside Barcelona or in areas with higher employment rates or a lower percentage of employment in services exhibit themselves higher employment rates.

[Table 5 here]

Notice also that the correlation coefficient is always positive and significantly different from zero, showing that the employment equation is not independent from the selection equation and that knowing Catalan has a positive effect in employment. I report likelihood ratio tests that also reject the hypothesis of independence between the language and employment equation, implying that for estimating the employment equations it is important to correct for language selection.

[Table 6 here]

Table 6 shows the predicted probabilities of being employed by Catalan reading and speaking and writing skills based on the previous estimations. Notice that predictions are very close to the actual values. This table also reports counterfactual probabilities: the probability of being employed for individuals who do not know Catalan were they to know it, and the probability of being employed for individuals who know Catalan were they not to know it. With the counterfactuals, I compute the premium of knowing Catalan, that is, its contribution to being employed for the different subsamples.[16] I also compute the counterfactual premia resulting from switching attributes by gender and by year.

The premium is positive and significant for all subsamples: for reading and speaking Catalan it is around 3% for men and in between 4 and 5% for women; for writing

it is around 2% for men and in between 5 and 6% for women. Notice that these premia are smaller than those one would obtain by a descriptive direct comparison across samples (Table 1), but generally greater than what one would estimate using a simple probit model (Table 2).

According to the theory of comparative advantage, individuals with positive economic rewards for knowing Catalan will actually know Catalan. This indicates that the economic incentive is not enough to explain individuals' decisions on language knowledge. Non-economic factors such as taste for language are playing an important role in language selection. On the other hand, one can also argue that this premium is measured in a transitional phase and individuals who do not know Catalan have more economic incentives to learn it. One would then expect the premium measured at a later stage of the Normalization process to conform more closely to the theory of comparative advantage. Evidence that this may already be happening is that while in 1991 individuals who do not know Catalan exhibit a larger premium than those who know it, in 1996 the premium for reading and speaking Catalan is larger for those individuals who know Catalan.

In terms of gender, for men the premium is larger for speaking and reading than for writing Catalan; for women it is larger for writing Catalan than for reading and speaking it. The premium is also larger for women than for men. However, from 1991 to 1996 the premium decreases for women, while it tends to increase for men, especially in reading and writing.

I next discuss the sensitivity of the premium to the instruments used, following Altonji, Elder and Taber (2002). This analysis is basically heuristic account taken of the fact that models of limited dependent variables do not have estimated sample errors and, therefore, do not have established tests of instruments, as do linear models.

[Table 7 here]

Table 7 presents the premium for all subsamples by exclusion of instruments, [17] that is, for the simple probit estimation, the baseline bivariate probit estimation, and the estimations without Municipal variables (percentage of people born in Catalonia and percentage of people who write Catalan in the *municipi*), and without origin variables (*not born in Catalonia*, *arrived younger than 10*, *years since migration*, *arrived younger than 10*$\times$*years since migration*, *born in Andalusia*, *Valencia-Balearics*, *Franja*).

Between a simple probit model and the baseline bivariate probit model there is an average difference of up to 0.7 percentage points, that is, 30% in relative terms. Correcting for selection into language knowledge seems to be important for a correct computation of the language premium.

On the other hand, excluding Municipal variables reduces the premium in 1991, but increases it in 1996. On the contrary, excluding origin variables increases the premium for all samples, except for men who read and speak Catalan in 1996, and especially for individuals who do not know Catalan. Hence, the premium is shown to be sensitive to exclusion restrictions, which shows that identification of the premium does not come from the nonlinear functional form, but from the instruments used.

Additionally, in order to check whether it is the lower female participation rate that explains the premium being larger for women than for men. I also report in Table 7 the premium recalculated by adding non-participants to the sample and computing them as non-employed. This variation results in increased premia for all groups, especially for women. Therefore, the larger premium for women does not seem to be the result of selection into labor market participation.[18]

5 Conclusions

In this paper I find evidence that the returns to reading, speaking and writing Catalan are positive for men and women for census years 1991 and 1996. Accounting for the

selection into knowing Catalan, I find that speaking and reading Catalan increases the probability of being employed by 3 to 5 percentage points, whereas writing Catalan increases the probability of being employed in between 2 and 6 percentage points. The premium is larger for women than for men and positive both for individuals who know Catalan and for those who do not. In spite of this economic reward, I find that non-economic considerations are also very important in explaining Catalan proficiency. In particular, externalities, such as living in a community where people know Catalan and are Catalan born, increase the probability that an individual knows Catalan.

These results should encourage research on multilingual labor markets, even if labor data are very scarce and the researcher is forced to rely on census data. Nevertheless, two caveats should be taken into consideration for improvement in further research on this subject. First, the estimated language premium is computed using Municipal, regional origin (including *Years Since Migration*) and Normalization variables as instruments, which is in agreement with the specificities of the Catalan scenario. However, the reader should also take these results cautiously, as the assumptions underlying the identification of the premium in this context may not be extendable to other contexts; on the other hand, these results are subject to improvement as richer datasets and wider sets of instruments become available for future research. Secondly, in this paper returns to language are measured as a difference of employment probabilities, yet they are found to be significant and show sensible trends. Measurement of these returns in terms of wages, as it is done in other contexts, would be a fruitful extension, feasible when and if employment and wage surveys in Catalonia include questions on language proficiency.

Appendix

A.1. Sample Selection

The following table illustrates the importance of the selection criteria in constructing the sample.

Census year	1991	1996
Total sample	250 000	250 000
Only main household members: parents and children	-17 654	-17 903
Only individuals between 16 and 60 years old	-82 297	-81 770
Only Spaniards	-5 740	-4 745
Only if variable arrival in Catalonia available		-3 788
Only individuals in the labor force	-47 421	-44 809
Only if Catalan language variable available	-25	
Selected sample	96 863	96 985

A.2. Definition of the variables

Employment.- The census reports a variable called "relationship with the activity." Somebody is employed if he or she reports to be working; somebody is unemployed, if he or she reports to be not working and looking for his/her first job or having worked before. Individuals who report other options (military service, retired, student, working at home) are not in the sample.

Schooling.- The census reports the maximum level of studies attained by the individual. To each level, I assign the number of years of schooling.

Age.- It is the census year, 1991 or 1996, respectively, minus the year of birth.

Normalization.- If the individual was younger than 12 in 1984, this dummy variable takes the value of one; it is zero otherwise.

Married.- This variable is one, if the respondent reports to be currently married; it is zero if s/he reports to be a widow(er), separated, or divorced.

Residence variables.- The census reports the *municipi* and the Province of residence for each individual. With this information I construct dummies for Lleida, Girona and Tarragona.

Origin variables.-The census reports the *municipi* and the province of birth for each individual. With this information I construct dummies for people who are not born in Catalonia, and born Andalusia, Valencia, Balearics and La Franja.

Years since Migration.- Census year minus the year of arrival to Catalonia. I also construct a dummy variable indicating if somebody arrived when s/he was younger than 10 years.

Municipal variables.- I use the residence variable to assign to each individual the corresponding information of the *municipi*.

Notes

[1]Castilian is generally known as "Spanish" outside the Iberian Peninsula, except in parts of Latin-America.

[2]These are the three main areas where Catalan is spoken. It should be noted that in Valencia census data refer to "Valencian," which is generally recognized as Catalan in Catalonia, but not always in Valencia. Figure 1 is based on the data presented in Table 1, pp 130-131, of Vallverdú (1990). I updated the percentages for 1991 using census data, which are available for the three areas. For 1996 there are data only for Catalonia.

[3]These three languages are relatively close to each other, but distant enough to prevent people from communicating if the speaker of one language does not know the other.

[4]The interested reader will find a chronology of some language-related historical facts, a description of Normalization measures, and the articles referring to language of the Spanish Constitution and of the Autonomy Statute for Catalonia in Rendon (2003, Tables 1 and 2 and Appendix A.1).

[5]The language of instruction in Catalan universities is not clearly determined by law, leaving it to the instructor's choice.

[6]Unfortunately, in this sample there is no way to relate individuals across time or to use family level information, that is, to link two individuals of the same family. Having infomation on the parents' language proficiency is necessary to study intergenerational language persistence or to restrict the sample to children of non-Catalan speakers, for whom learning Catalan means more of an investment. Consequently, these data only allow us to analyze the cross-sectional effect of individual attributes on language

selection.

[7]In 1991 the international immigration rate in Catalonia was very low, 4%.

[8]Unfortunately, as Charette and Meng (1994) point out, self-assessed data can lead to biased inference. Corrections for missclassification, such as the one used by Dustmann and van Soest (2001), require panel data and are not possible in the current framework.

[9]In several estimations I found origin to be non-significant in determining employment status. However, origin may be more important explaining differences in wages, for which, unfortunately, there are no data in the census.

[10]In the census, the lowest level of schooling is to be illiterate, followed by 'no schooling.' The percentage of illiterates in the sample is negligible, while the percentage with no schooling is very small.

[11]For a model that incorporates marriage markets, language choice, and earnings outcomes see Robinson (1988).

[12]I also estimated a probit with fixed effects for the *municipis* and an ordered probit with three categories for each profficiency level: basic level, reading and speaking, and writing Catalan. Those results are essentially the same as these shown here.

[13]In this cross-section it is impossible to separate the cohort effect from the years since migration effect. See for instance Borjas (1994).

[14]It may very well be that their jobs and wages are different from those of individuals born in Catalonia, but to investigate this distinction would require another dataset and it is beyond the purpose of this paper.

[15]This is another feature that makes this scenario so particular: not all locals *know* their language, but all locals know the *immigrants*' language.

[16]Direct and indirect marginal effects of the covariates on the premium computed in a similar estimation are presented in Rendon (2003).

[17]Variations in the premium across specifications stem basically from variations in the counterfactual employment probability. The predicted employment probability is close to the actual probability and does not change more than 0.01% across specifications. It is thus enough to report the premium.

[18]Explaining why the premium is larger for women than for men would require extending the analysis to employment by sectors or occupations, which is beyond the scope of the current paper.

References

Altonji, J. G., Elder, T. E. and Taber, C. R. (2002), An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schools. NBER WP. No. 9358.

Angrist, J. D. and Lavy, V. (1997), 'The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco', *Journal of Labor Economics* **15**, 48–76.

Borjas, G. (1987), 'Self-Selection and the Earnings of Immigrants', *American Economic Review* **77**(27), 21–37.

Borjas, G. (1995), 'Assimilation and Changes in Cohort Quality Revisited: What happened to Immigrant Earnings in the 1980s', *Journal of Labor Economics* **13**, 201–45.

Cattaneo, A. and Winkelmann, R. (2005), 'Earnings Differentials between German and French Speakers in Switzerland', *Revue Suisse d'Economie et de Statistique* **2**, 191–212.

Charette, M. and Meng, R. (1994), 'Explaining language proficiency. Objective versus self-assessed measures of literacy', *Economic Letters* **44**, 313–321.

Chiswick, B. R. and Miller, P. W. (1995), 'The Endogeneity Between Language and Earnings: International Analyses', *Journal of Labor Economics* **13**, 246–88.

Chiswick, B. R. and Miller, P. W. (2002), 'Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle.', *Journal of Population Economics* **15**, 31–57.

Christofides, L. N. and Swidinsky, R. (2002), Language and Earnings in the Federal Public Service in Canada. University of Guelph. Department of Economics. Mimeo.

Clark, K. and Drinkwater, S. (2002), 'Enclaves, Neighbourhood Effects and Employment Outcomes: Ethnic Minorities in England and Wales.', *Journal of Population Economics* **15**, 5–29.

Drinkwater, S. J. and O'Leary, N. C. (1997), 'Unemployment in Wales: Does Language Matter?', *Regional Studies* **31**, 581–591.

Dustmann, C. (1994), 'Speaking Fluency, Writing Fluency and Earnings of Migrants', *Journal of Population Economics* **7**, 133–156.

Dustmann, C. and Fabbri, F. (2003), 'Language Proficiency and Labour Market Performance of Immigrants in the UK', *Economic Journal* **113**, 695–717.

Dustmann, C. and van Soest, A. (2001), 'Language Fluency and Earnings: Estimation with Misclassified Language Indicators', *Review of Economics and Statistics* **83**, 663–674.

Henley, A. and Jones, R. E. (2003), Earnings and Linguistic Ability in a Bilingual Economy. 106. Royal Economic Society Annual Conference.

Kossoudji, S. (1988), 'English Language Ability and the Labor Markets Opportunities of Hispanic and East Asian immigrant men', *Journal of Labor Economics* **6**(2), 205–228.

Lang, K. (1986), 'A Language Theory of Discrimination', *Quarterly Journal of Economics* **101**(2), 363–382.

Lazear, E. (1999), 'Culture and Language', *Journal of Political Economy* **107**(6), S96–S126.

Maddala, G. S. (1983), *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.

McManus, W. S. (1985), 'Labor Market Costs of Language Disparity: An Interpretation of Hispanic Earnings Differences', *American Economic Review* **75**(4), 818–827.

Rendon, S. (2003), The Catalan Premium: Work and Language in Catalonia. Universidad Carlos III de Madrid. Working Paper No. 03-34.
URL: *http://docubib.uc3m.es/WORKINGPAPERS/WE/we033410.pdf*

Rivera-Batiz, F. (1990), 'English language proficiency and the economic progress of immigrants', *Economic Letters* **34**, 295–300.

Robinson, C. (1988), Language Choice: the Distribution of Language Skills and Earnings in a Dual-Language Economy, *in* E. R. Ehrenberg, ed., 'Research in Labor Economics', JAI Press Inc., London, pp. 53–90.

Saarela, J. and Finnäs, F. (2003), 'Unemployment and native language: the Finnish case', *The Journal of Socio-Economics* **32**, 59–80.

Strubell i Trueta, M. (1981), *Llengua i Població a Catalunya*, Edicions de la Magrana, Barcelona.

Tudela, J. (1986), *El futur del català. Una radiografia sociolingüstica*, Publicacions El Món S.A., Barcelona.

Vallverdú, F. (1970), *Dues llengües: dues funcions? Per una històrica lingüstica de la Catalunya contemporània.*, Edicions 62, Barcelona.

Vallverdú, F. (1990), *L'ús del català: Un futur controvertit*, Edicions 62, Barcelona.

Willis, R. and Rosen, S. (1979), 'Education and Self-Selection', *Journal of Political Economy* **87**(5), S7–S36.

Woolard, K. A. (1989), *Double Talk. Bilingualism and the politics of ethnicity in Catalonia*, Stanford University Press cop., Stanford, California.

Table 1: Summary Statistics

Census Year	1991		1996	
Gender	Men	Wom.	Men	Wom.
Catalan Knowledge				
% Basic level	37.7	29.5	26.3	18.5
% Speak & Read	25.8	22.9	27.5	22.4
% Write	36.5	47.6	46.2	59.1
Catalan Knowledge of those not born in Catalonia				
% Basic level	69.8	62.5	59.3	49.6
% Speak & Read	19.9	21.2	26.8	27.0
% Write	10.4	16.3	13.9	23.3
Employment				
%Employed	89.9	78.5	85.1	74.5
%Employed of those not born in Catalonia	91.0	77.0	86.3	74.2
%Employed by Catalan Knowledge				
Basic level	87.8	72.3	81.9	68.0
Speak & Read	92.0	78.5	87.3	73.3
Write	91.0	82.5	85.6	77.0
Schooling	7.7	8.3	8.6	9.2
% Illiterate	0.4	0.5	0.3	0.3
% No schooling	8.5	6.1	6.1	4.6
Age	38.0	34.4	38.2	35.5
% affected by *normalització*	3.9	5.0	13.0	15.9
% Married	66.6	58.2	62.8	57.4
% Lleida residents	5.7	5.0	5.9	5.2
% Girona residents	8.4	8.4	8.5	8.7
% Tarragona residents	8.9	7.7	9.4	8.5
% employed over population in *municipi*	37.2	37.4	36.3	36.3
% work in Services in *municipi*	51.7	52.5	57.4	58.2
% writes in *municipi*	39.7	40.1	45.6	45.9
% Catalan-born in *municipi*	67.1	67.2	68.2	68.1
% not born in Catalonia	37.6	30.2	31.0	24.9
% arrived age ≤ 9	11.0	11.1	10.0	9.6
Years since migration if not born in Cat	25.0	23.6	28.4	26.8
% born in Andalusia	24.1	18.3	19.5	14.4
% born in Valencia	1.1	1.1	1.0	0.9
% born in Balearics	0.1	0.1	0.1	0.2
% born in Franja	0.5	0.4	0.4	0.4
Number of observations	60446	36417	57785	39200

Table 2: Effect of Catalan Knowledge on Employment Status. Probit estimations with a Dummy variable for Catalan Skill. Standard errors in small fonts

Census Year	1991				1996			
Gender	Men		Women		Men		Women	
Knowledge	Knw.	D.knw	Knw.	D.knw	Knw.	D.knw	Knw.	D.knw
Reading and Speaking	1.96	2.14	4.68	5.22	2.83	3.09	4.33	4.71
	1.34	1.25	1.78	1.51	1.67	1.48	1.65	1.37
Writing	1.82	1.74	5.91	6.25	2.35	2.29	5.82	6.00
	1.32	1.15	2.23	1.90	1.49	1.30	2.00	1.67

Table 3: Catalan Knowledge Equation. Probit estimations. Standard errors in small fonts

Catalan Skill	Reading and Speaking				Writing			
Census Year	1991		1996		1991		1996	
Gender	Men	Wom.	Men	Wom.	Men	Wom.	Men	Wom.
Constant	-3.98	-4.16	-4.17	-4.08	-1.56	-1.49	-0.73	-1.30
	0.17	0.24	0.22	0.30	0.16	0.21	0.19	0.24
Schooling $\times 10^{-1}$	3.54	3.85	4.39	4.81	3.75	4.13	4.42	4.84
	0.10	0.15	0.15	0.20	0.10	0.13	0.13	0.18
Schooling^2 $\times 10^{-3}$	-7.64	-9.02	-9.60	-11.36	-6.75	-8.76	-7.01	-9.70
	0.34	0.49	0.57	0.77	0.31	0.44	0.55	0.70
Age $\times 10^{-2}$	6.19	6.69	2.84	3.03	-3.68	-4.33	-10.26	-9.35
	0.53	0.73	0.70	0.96	0.50	0.65	0.62	0.80
Age^2 $\times 10^{-4}$	-3.96	-4.60	0.05	-0.34	6.91	7.18	13.99	12.29
	0.59	0.83	0.78	1.09	0.57	0.76	0.71	0.92
Age×Schooling $\times 10^{-1}$	-2.12	-2.00	-2.66	-2.56	-2.73	-2.08	-3.56	-2.72
	0.18	0.27	0.23	0.32	0.17	0.24	0.20	0.26
Married $\times 10^{-3}$	-8.76	-120.37	13.86	-83.05	-196.58	-202.38	-120.14	-182.60
	17.09	19.87	17.98	21.12	16.30	18.17	15.79	17.56
Lleida resident $\times 10^{-1}$	2.63	1.69	1.15	0.91	1.12	1.65	1.39	1.21
	0.33	0.47	0.37	0.54	0.29	0.39	0.29	0.39
Girona resident $\times 10^{-1}$	2.35	2.42	2.39	1.63	1.68	1.98	2.68	1.46
	0.26	0.36	0.31	0.42	0.23	0.31	0.25	0.31
Tarragona resident $\times 10^{-1}$	1.62	2.10	2.16	1.48	1.28	1.34	1.83	1.45
	0.24	0.34	0.25	0.35	0.23	0.31	0.22	0.28
% Mun[a] Services $\times 10^{-1}$	-7.75	-7.31	-3.78	-4.55	-9.76	-10.36	-9.69	-7.86
	0.57	0.76	0.68	0.89	0.54	0.69	0.59	0.73
% Mun[a] Write Catalan	2.71	2.41	2.74	2.30	3.90	3.64	4.01	3.48
	0.11	0.16	0.14	0.21	0.10	0.14	0.12	0.16
% Mun[a] Employed $\times 10^{-1}$	7.43	7.85	7.67	5.42	-7.75	-12.64	-9.74	-5.08
	3.68	5.13	3.46	4.69	3.35	4.55	2.92	3.79
% Mun[a] Born in Catalonia $\times 10^{-1}$	8.55	11.37	12.43	15.56	-8.25	-4.72	-5.46	-1.25
	1.11	1.64	1.33	1.96	1.03	1.43	1.10	1.49
Normalized[b] $\times 10^{-1}$	3.48	3.00	3.10	2.96	1.81	1.66	0.03	1.05
	0.38	0.45	0.36	0.46	0.35	0.41	0.30	0.37
Not born in Cat	-1.86	-1.71	-1.99	-1.77	-1.40	-1.34	-1.57	-1.39
	0.04	0.05	0.04	0.05	0.05	0.06	0.05	0.06
Arrived younger than 10 $\times 10^{-1}$	8.14	8.41	9.83	9.93	4.49	5.39	8.36	8.37
	0.65	0.84	0.79	1.03	0.75	0.93	0.86	1.04
Years Since Migration$\times 10^{-2}$	1.56	1.46	1.94	1.60	0.88	1.03	1.40	1.32
	0.13	0.19	0.14	0.19	0.18	0.24	0.18	0.22
Arrived younger than 10 $\times$YSM[c] $\times 10^{-3}$	-7.35	-9.28	-11.23	-13.91	-2.47	-6.49	-11.77	-14.04
	2.13	2.99	2.34	3.19	2.58	3.49	2.66	3.40
Born in Andalusia $\times 10^{-2}$	-1.45	-10.58	-6.10	-8.16	-1.28	-5.58	-3.79	-9.69
	2.11	2.93	2.30	3.08	2.80	3.57	2.95	3.53
Born in Valencia -Balearics$\times 10^{-1}$	2.31	1.59	1.87	3.39	2.86	0.62	2.49	1.54
	0.55	0.75	0.64	0.87	0.61	0.79	0.64	0.79
Born in Franja $\times 10^{-1}$	3.04	5.83	2.44	3.80	3.02	3.46	1.54	0.74
	0.85	1.27	1.02	1.41	0.91	1.17	1.03	1.21
Pseudo R^2	0.33	0.32	0.35	0.35	0.29	0.30	0.31	0.33

[a] Percentage in *municipi*; [b] Affected by *normalització*; [c] YSM=Years Since Migration.

Table 4: Predicted Probability of knowing Catalan (in percent) Standard errors in small fonts

Catalan Skill	Reading and Speaking				Writing			
Census Year	1991		1996		1991		1996	
Gender	Men	Women	Men	Women	Men	Women	Men	Women
Actual	62.31	70.54	73.72	81.54	36.51	47.63	46.27	59.10
Predicted	62.36	70.57	73.79	81.59	36.57	47.70	46.33	59.16
	0.67	0.15	0.65	0.12	0.40	0.16	0.42	0.15
Counterfactuals								
Switch year	62.30	72.52	73.19	79.82	36.36	49.87	47.64	58.29
	0.67	0.15	0.65	0.12	0.40	0.17	0.43	0.14
Diff. year	-0.06	1.95	-0.60	-1.77	-0.21	2.17	1.31	-0.87
	0.95	0.21	0.92	0.17	0.57	0.23	0.60	0.21
Switch gender	65.68	67.14	77.50	78.10	41.05	42.95	52.48	53.03
	0.70	0.15	0.69	0.13	0.45	0.15	0.48	0.15
Diff. gender	3.32	-3.44	3.71	-3.49	4.48	-4.74	6.15	-6.12
	0.97	0.21	0.95	0.17	0.60	0.21	0.64	0.22

Table 5. Employment Equation by Reading and Speaking, and Writing Skills
Standard errors in small fonts

Catalan Skill	Reading and Speaking				Writing			
Census Year	1991		1996		1991		1996	
Gender	Men	Wom.	Men	Wom.	Men	Wom.	Men	Wom.
Constant	-2.56	-3.02	-3.34	-2.28	-2.56	-3.02	-3.34	-2.28
	0.17	0.19	0.16	0.18	0.17	0.19	0.16	0.18
Schooling	7.90	6.91	10.32	8.12	7.93	6.95	10.29	8.04
$\times 10^{-2}$	1.15	1.41	1.35	1.58	1.15	1.41	1.35	1.58
Schooling^2	-22.92	3.40	-24.01	-4.27	-23.12	2.87	-23.95	-4.11
$\times 10^{-4}$	4.07	4.79	5.14	5.61	4.06	4.77	5.14	5.61
Age	1.01	0.64	1.19	0.62	1.01	0.64	1.19	0.62
$\times 10^{-1}$	0.05	0.06	0.05	0.06	0.05	0.06	0.05	0.06
Age ^{2}	-1.24	-0.52	-1.44	-0.52	-1.23	-0.52	-1.43	-0.52
$\times 10^{-3}$	0.06	0.07	0.05	0.07	0.06	0.07	0.05	0.07
Age $\times$Schooling	3.07	-1.93	0.68	1.15	3.14	-1.76	0.73	1.31
$\times 10^{-4}$	1.98	2.52	1.93	2.34	1.99	2.52	1.93	2.34
Married	0.51	-0.05	4.69	0.36	5.11	-0.51	4.70	0.37
$\times 10^{-1}$	0.19	0.18	0.17	0.17	0.19	0.18	0.17	0.17
Lleida resident	3.04	3.24	3.40	2.39	3.02	3.24	3.40	2.38
$\times 10^{-1}$	0.37	0.39	0.34	0.35	0.37	0.39	0.34	0.35
Girona resident	2.16	0.95	2.06	2.64	2.16	0.95	2.07	2.63
$\times 10^{-1}$	0.31	0.30	0.30	0.30	0.31	0.30	0.30	0.30
Tarragona resident	2.08	2.08	1.72	1.82	2.09	2.09	1.72	1.82
$\times 10^{-1}$	0.27	0.29	0.24	0.26	0.28	0.29	0.24	0.26
% Mun^a Employed	4.01	5.30	4.36	2.64	4.04	5.29	4.37	2.65
	0.38	0.39	0.30	0.31	0.38	0.39	0.30	0.31
% Mun^a Services	-6.77	-3.63	-7.40	-4.86	-6.79	-3.65	-7.39	-4.89
$\times 10^{-1}$	0.66	0.67	0.63	0.65	0.66	0.67	0.63	0.64
ρ	0.08	0.10	0.08	0.06	0.07	0.12	0.05	0.09
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LRT ($\rho = 0$)	61.85	70.18	57.47	27.35	35.84	113.38	22.07	75.97

[a] Percentage in *municipi.*

Table 6: Probability of employment and Premium (in percent) by Catalan skills
Standard errors in small fonts

Census Year	1991				1996			
Gender	Men		Women		Men		Women	
Knowledge	Knw.	D.knw	Knw.	D.knw	Knw.	D.knw	Knw.	D.knw
READING AND SPEAKING								
Actual	91.05	87.97	81.16	72.28	86.21	81.91	75.99	68.04
Pred: Know	91.00	90.62	81.54	79.51	86.03	85.28	75.72	72.80
	7.30	7.13	8.23	7.40	10.59	9.93	11.17	9.65
Pred: DKnow	88.49	88.06	76.59	74.40	82.64	82.07	71.58	69.11
	8.54	8.27	9.12	8.26	11.96	11.10	11.82	10.22
Premium	2.51	2.57	4.95	5.11	3.40	3.21	4.14	3.69
	1.30	1.23	1.12	1.17	1.54	1.31	1.02	0.78
Counterfactual Premium:								
Switch year	3.05	3.24	3.81	3.92	2.83	2.61	5.42	4.98
	1.41	1.34	0.87	0.90	1.44	1.24	1.27	1.10
Switch gender	4.88	4.95	2.49	2.56	3.97	3.59	3.53	3.27
	1.23	1.17	1.25	1.17	1.02	0.81	1.59	1.32
WRITING								
Actual	90.40	89.59	82.45	74.99	85.57	84.66	77.02	70.91
Pred: Know	90.28	91.64	82.65	83.01	85.20	86.68	76.51	76.85
	7.94	6.78	8.74	7.12	11.35	9.62	11.59	9.54
Pred: DKnow	88.44	89.66	77.29	76.52	83.43	84.81	71.36	71.59
	8.96	7.68	10.16	8.35	12.18	10.35	12.75	10.45
Premium	1.84	1.98	5.37	6.49	1.77	1.87	5.15	5.26
	1.06	1.05	1.58	2.08	0.91	0.91	1.50	1.50
Counterfactual Premium:								
Switch year	1.68	1.99	4.96	6.10	1.98	1.92	5.70	5.87
	0.89	1.00	1.42	2.06	1.12	1.01	1.64	1.62
Switch gender	5.28	6.10	1.81	2.05	4.98	5.13	1.81	1.89
	1.68	1.99	1.03	1.07	1.51	1.56	0.91	0.91

Table 7 Premium (in %) computed excluding municipal and origin variables, and including non labor force participants. Standard errors in small fonts

Group	Men 1991		Women 1991		Men 1996		Women 1996	
Knowledge	Knw.	D.Knw	Knw.	D.Knw	Knw.	D.Knw	Knw.	D.Knw
READING AND SPEAKING								
Simple Probit	1.96	2.14	4.68	5.22	2.83	3.09	4.33	4.71
	1.34	1.25	1.78	1.51	1.67	1.48	1.65	1.37
Baseline Bivariate	2.51	2.57	4.95	5.11	3.40	3.21	4.14	3.69
	1.30	1.23	1.12	1.17	1.54	1.31	1.02	0.78
No *municipi*	2.35	2.44	4.88	5.00	3.43	3.29	4.43	4.01
	1.20	1.16	1.06	1.09	1.53	1.33	1.01	0.83
No origin	2.63	3.39	5.60	7.53	3.17	4.29	4.33	5.97
	1.62	1.81	2.13	2.29	1.67	2.44	1.47	2.03
Plus non-participants	2.58	2.65	9.37	9.08	3.88	3.62	4.57	4.06
	1.30	1.23	1.98	1.54	1.68	1.43	1.06	0.82
WRITING								
Simple Probit	1.82	1.74	5.91	6.25	2.35	2.29	5.82	6.00
	1.32	1.15	2.23	1.90	1.49	1.30	2.00	1.67
Baseline Bivariate	1.84	1.98	5.37	6.49	1.77	1.87	5.15	5.26
	1.06	1.05	1.58	2.08	0.91	0.91	1.50	1.50
No *municipi*	1.76	1.90	5.27	6.05	1.90	2.01	5.29	5.43
	1.01	1.00	1.51	1.78	0.96	0.96	1.48	1.49
No origin	2.28	2.68	6.47	8.68	2.29	2.98	5.63	7.88
	1.32	1.45	2.45	2.87	1.16	3.20	1.93	6.45
Plus non-participants	1.85	2.02	8.06	9.89	2.05	2.15	5.35	5.46
	1.05	1.05	1.58	2.28	1.01	1.02	1.49	1.52

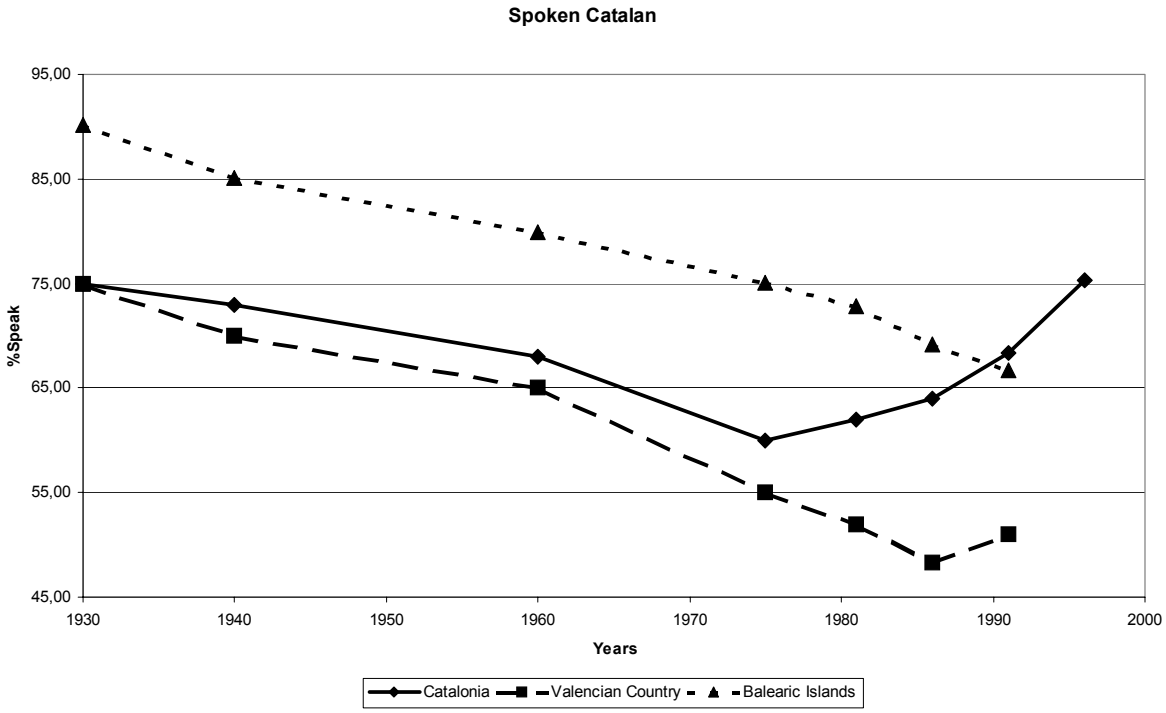

Figure 1: Spoken Catalan in Catalonia, Valencia and Balearics (Source: Vallverdú 1990)